Uncertainties of the 50-year wind from short time series using generalized extreme value distribution and generalized Pareto distribution

Larsén, Xiaoli Guo; Mann, Jakob; Rathmann, Ole; Ejsing Jørgensen, Hans

Published in:
Wind Energy

Link to article, DOI:
10.1002/we.1683

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Wind Energy

WIND ENERGY
Wind Energ. (2013)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/we.1683

RESEARCH ARTICLE

Uncertainties of the 50-year wind from short time series using generalized extreme value distribution and generalized Pareto distribution†

Xiaoli Guo Larsén, Jakob Mann, Ole Rathmann and Hans E. Jørgensen

Department of Wind Energy, Risø Campus, Technical University of Denmark, 4000 Roskilde, Denmark

ABSTRACT

This study examines the various sources to the uncertainties in the application of two widely used extreme value distribution functions, the generalized extreme value distribution (GEVD) and the generalized Pareto distribution (GPD). The study is done through the analysis of measurements from several Danish sites, where the extreme winds are caused by the Atlantic lows. The simple extreme wind mechanism here helps us to focus on the issues mostly related to the use of limited wind measurements. Warnings are flagged and possible solutions are discussed. Thus, this paper can be used as a guideline for applying GEVD and GPD to wind time series of limited length. The data analysis shows that, with reasonable choice of relevant parameters, GEVD and GPD give consistent estimates of the return winds. For GEVD, the base period should be chosen in accordance with the occurrence of the extreme wind events of the same mechanism. For GPD, the choices of the threshold, the definition of independent samples and the shape factor are interrelated. It is demonstrated that the lack of climatological representativity is a major source of uncertainty to the use of both GEVD and GPD; the information of climatological variability is suggested to be extracted from global or mesoscale models. © 2013 The Authors. *Wind Energy* published by John Wiley & Sons, Ltd.

KEYWORDS

the 50-year wind; generalized extreme value distribution; generalized Pareto distribution; uncertainty

Correspondence

Xiaoli Guo Larsén, Department of Wind Energy, Risø Campus, Technical University of Denmark, 4000 Roskilde, Denmark.
E-mail: xgal@dtu.dk

Received 30 May 2013; Revised 27 September 2013; Accepted 27 September 2013

1. INTRODUCTION

In wind energy industry, the turbine design wind needs to be estimated in order to ensure that winds will not exceed the turbine's design specification and to avoid turbine design being unrealistically over-specified. This design wind speed is usually defined as the 50-year wind. In the current European load code,[1] the 50-year wind is referred to as the 10 min averaged wind that, on average, occurs once in 50 years. One of the biggest challenges in estimating the 50-year wind is the shortage of long-term wind measurements. To confront this issue, statistical and numerical models have been used to generate long-term artificial data. For instance, the Monte Carlo method is often used to generate hundreds or even thousands of years of data[2] on the basis of a short time series. However, these artificial data are questionable regarding their representativity for the natural climatological variability in strong winds, which is crucial for deducing the distribution of the extreme wind samples. Such climatological information can sometimes be obtained from long term measurements, such as those conventionally from masts, but also from altimeters and buoys.[3,4] However, measurements are often challenged with insufficient temporal coverage and resolution, limited spatial availability and inconsistency in quality through time, all of which are important for assessing the strong wind climatology. It is generally considered that numerically modeled climate data (e.g., hindcast) can reasonably provide such a climatological variability, but these data are troubled with the smoothing effect of models that leads to a lack of variability at strong winds, which is also crucial for the extreme wind estimation.[5]

Table I. List of symbols and abbreviations.

Variables	Definition
A	A scale factor as in Equation (9), ms^{-1}
D_c	95% Confidence interval, $1.36/\sqrt{n}$, for goodness-of-fit for F
E	Observed mean exceedance, ms^{-1}
$F(U)$	The distribution function for the extreme winds
$\max(\Delta F)$	The maximum difference between the observed and fitted $F(U)$
k	A shape factor deciding the type of distribution
k_T	A frequency factor, equation (8)
L	Wind data length in years
n	Number of extreme wind samples
N	Record length in years
T	Return period
T_{BP}	Base period for PMM
Δt	Minimum time interval used to select independent storm events
u_0	Threshold, ms^{-1}
u_{01}	The first value of the chosen threshold list for obtaining k and A, ms^{-1}
u_{0h}	Threshold where $\lambda \sim 1$, ms^{-1}
U	Extreme wind values, ms^{-1}
U^{max}	List of annual wind maxima, ms^{-1}
U_T	T-year return wind, ms^{-1}
α	Distribution parameter
β	Distribution parameter, ms^{-1}
λ	Exceedance rate, per year
λ_0	Exceedance rate, per year, at wind speed threshold u_0
σ	Standard deviation of U^{max}, $\pi/(\alpha\sqrt{6})$, ms^{-1}
$\sigma(U_T)$	Standard error, equation (7), a function of σ and k_T, ms^{-1}
γ_E	The Eular constant

Abbreviation	Definition
GEVD	Generalized extreme value distribution
GPD	Generalized Pareto distribution
PMM	Periodical maximum method
AMM	Annual maximum method
POT	Peak-over-threshold method
GOF	Goodness-of-fit

These problems related to the modeling have partly led people to remain solely with measurements, regardless of how short the time series is. For estimating the extreme winds, the generalized extreme value distribution (GEVD) and the generalized Pareto distribution (GPD) are the most used distribution functions. For industrial use, in order to compensate for the problems with insufficient extreme wind samples caused by short data series length, it is popular to lower the wind speed threshold for more exceedances in connection with the use of GPD or to shorten the basis period in order to get more periodic wind maxima in connection with the use of GEVD. These simple attempts can, however, easily violate the pre-conditions for applying these distribution functions, and it is therefore necessary to examine the extent to which the measurements of only a limited number of years can be used for estimating the 50 year wind. The current paper sets therefore the focus on investigating the uncertainties that are related to a number of factors in using GEVD and GPD, especially when using short time series.

These factors are examined through measurements from a number of Danish stations. One unique advantage of these data is that the corresponding extreme winds are of only one mechanism, namely the Atlantic lows. This simple character helps us avoid the complicated issues related to multimechanisms as in some other places[6] and helps us focus on issues related to the use of short time series.

The algorithms for the return wind using GEVD and GPD are presented in Section 2. Measurements are introduced in Section 3. The results on the uncertainty analysis are presented in Section 4, followed by discussions and conclusions in Sections 5 and 6. A list of variables and abbreviations is given in Table I for readability.

2. BACKGROUND

To understand the sources and impact of the uncertainties of the extreme wind estimation, it is necessary to have an overview of how the key parameters are calculated. The algorithms for GEVD and GPD are reviewed in the following.

2.1. GEVD

The generalized extreme value cumulative distribution for fitting the extreme wind values in the form of wind maxima from a basis period T_{BP} takes the form:

$$F(U) = \exp\left(-(1-\alpha k(U-\beta))^{1/k}\right) \tag{1}$$

where $F(U)$ is the probability that wind speed U is not exceeded during the basis period, k is a shape factor and α and β are distribution parameters. For $k > 0$, GEVD is known as a type III (or reverse Weibull) extreme value distribution. For $k < 0$, GEVD is known as a type II (or Frechet) extreme wind distribution. For $k = 0$, GEVD :

$$F(U) = \exp\left(-\exp(-\alpha(U-\beta))\right) \tag{2}$$

Note that equations (1) and (2) are the integration of the corresponding probability density functions for the extreme wind samples U, given that these samples are independent and identically distributed.

Because of its association to a certain basis period, T_{BP}, the method is denoted the periodic maximum method (PMM) or, in the case of a basis period of 1 year, annual maximum method (AMM).

Equating $1/(T/T_{BP})$ with $1 - F(U)$, with $F(U)$ as in equation (1), gives the T-year return wind U_T:

$$U_T = (\alpha k)^{-1}\left(1 - (\ln\frac{T/T_{BP}}{T/T_{BP}-1})^k\right) + \beta \tag{3}$$

Equating $1/(T/T_{BP})$ with $1 - F(U)$, with $F(U)$ as in equation (2), gives the T-year return wind U_T for the Gumbel distribution at a relatively long return period ($T \gg 1$) as

$$U_T = \alpha^{-1}\ln(T/T_{BP}) + \beta \tag{4}$$

where α and β can be obtained with a couple of methods. One method is applying least square regression to U_i^{max} (i=1, n), the ranked sequence of wind maxima from n segments of the entire time series, versus $-\ln[-\ln(i/(n+1))]$. Another method is the probability-weighted moment procedure:[7,8]

$$\alpha = \frac{\ln 2}{2b_1 - \overline{U^{max}}}, \quad \beta = \overline{U^{max}} - \frac{\gamma_E}{\alpha} \tag{5}$$

where $\gamma_E \approx 0.577215665$ is Euler's constant, and $\overline{U^{max}}$ is the mean of U_i^{max}. b_1 is calculated from

$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{i-1}{n-1}U_i^{max} \tag{6}$$

The values of α and β are similar using the two methods. According to Abild and Hosking,[7,8] the probability-weighted moment procedure yields less bias and variance on the parameter estimates and has been proven highly efficient even for small size samples, and it is used here for obtaining the 50-year wind.

Figure 1 illustrates the curvature of U_T versus T for the three types of distribution using AMM. Both types I and II give unbounded high values at very large T, which reflects the non-physical characteristics when regarding the extreme events as wind. Type II, especially, is highly uncertain for extrapolation of the extreme winds. Type III corresponds to a return wind that approaches a limiting value at high return period T.

Mann *et al.*[9] gave the standard error of the Gumbel fitting in obtaining U_T from the standard deviation of U^{max} as

$$\sigma(U_T) = \frac{\pi}{\alpha}\sqrt{\frac{1 + 1.14k_T + 1.10k_T^2}{6n}} \tag{7}$$

where

$$k_T = -\frac{\sqrt{6}}{\pi}\left(\ln\ln\left(\frac{T}{T-1}\right) + \gamma_E\right) \tag{8}$$

Kite[10] showed that the T-year estimate can be considered as normally distributed, and accordingly, the 95% confidence interval can be calculated as $U_T \pm 1.96 \cdot \sigma(U_T)$ ([7]).

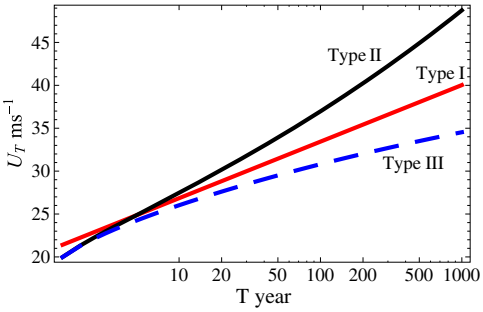

Figure 1. Illustration of the curvatures in log-linear scale related to the distribution of the T-year return wind with T for $k = 0$ (type I), $k < 0$ (type II) and $k > 0$ (type III), respectively.

2.2. GPD

The GPD is used to describe the wind speed exceedances over a threshold, and it takes the form:

$$\begin{aligned} F(U, u_0) &= 1 - \left(1 - \frac{k(U-u_0)}{A}\right)^{1/k}, \quad k \neq 0 \\ &= 1 - \exp\left(\frac{U-u_0}{A}\right), \quad k = 0 \end{aligned} \tag{9}$$

where u_0 is the speed threshold, A is a scale factor and k is a shape factor.

A Poisson process has been found to be a suitable way to describe how the individual exceedances occur randomly in time, independent of each other.[7] This has resulted in the name peak-over-threshold (POT). In practice, this requires application of a 'time separation filter' to ensure that the extreme wind events are sufficiently separated in time, a procedure used by Cook,[11] where it is also called the method of 'independent storm'.

For the same wind climate, the shape factor k has in fact the same value as that of GEVD because of the mathematical relationship between PMM and POT distributions as shown in Appendix A.

Similar to GEVD, $k = 0$, $k < 0$ and $k > 0$ correspond to extreme wind distribution of types I, II and III, respectively. Again, type III is the only type of the three that provides limiting return winds at high T. The impact of k on the estimate of U_T is similar to that shown in Figure 1. If the exceedance rate of the level u_0 is λ per year, then the mean crossing rate of the level U_T is $\lambda(1 - F(U_T, u_0))$. Relating $\lambda(1 - F(U_T, u_0))$ to $1/T$, together with equation (9), gives

$$U_T = u_0 + A\frac{1 - (\lambda T)^{-k}}{k} \tag{10}$$

which, for $k = 0$, simplifies into

$$U_T = u_0 + A\ln(\lambda T) \tag{11}$$

which can be written in a form similar to equation (4):

$$U_T = A\ln T + B \tag{12}$$

with $B = u_0 + A\ln\lambda$.

Abild and Mann *et al.*[7,9] used the Poisson process and properties of the exponential distribution; together with the propagation of variance formula, they obtained the uncertainty in relation to equation (11):

$$\sigma(U_T) \approx \frac{A}{\sqrt{\lambda L}}\sqrt{1 + \ln^2(\lambda T)} \tag{13}$$

where L is the data length. Similar to GEVD, the 95% confidence interval can be obtained as $U_T \pm 1.96 \cdot \sigma(U_T)$, under the assumption that the estimate of U_T is normally distributed.[10]

Table II. Details of measurements.

Stations	Data period	Height (m)	Wind speed data coverage
Tystofte	1982–2010	39.3	96.7%
Sprogø	1977–1999	70.0	97.8%
Kegnæs	1991–2006	23.4	99.5%
Jylex	1982–2004	24.0	96.4%
Risø	1996–2009	76.6	99.4%

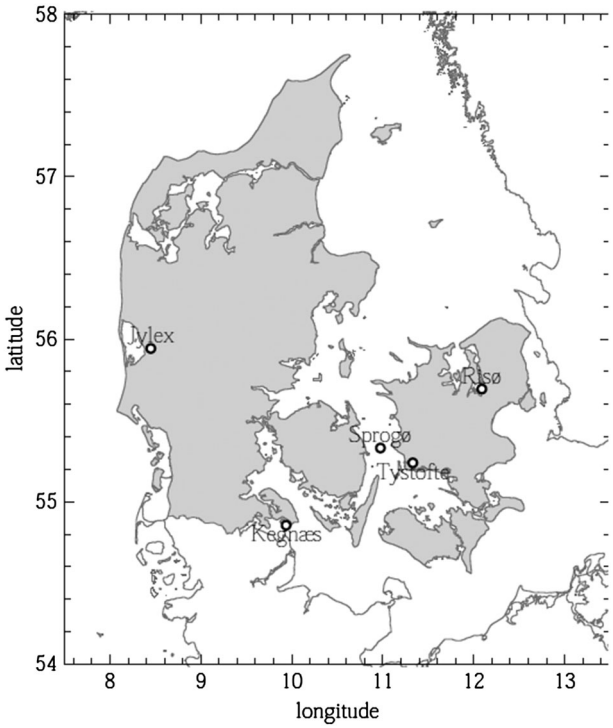

Figure 2. Map of Denmark and locations of the five sites.

3. MEASUREMENTS

Wind measurements from five Danish sites are used. The data are 10 min averages. Data period, measurement heights and average data coverage for each site are listed in Table II. The locations of these sites can be found in Figure 2. For these sites, the data record length varies from 14 to 29 years. These long-term measurements have provided a sound base for studying the various factors that cause the uncertainties in the estimation of return wind using short time series.

4. RESULTS: THE UNCERTAINTIES

One of the sources of uncertainty that is common in both GEVD and GPD is the determination of the shape factor k as in equations (1) and (9). The k-effect is examined in Section 4.1.

A number of other factors leading to uncertainties in using GEVD and GPD are studied in Sections 4.2 and 4.3, respectively.

4.1. The *k* effect

In obtaining the k value, it is risky to apply simple fitting of the samples using equations (1) or (9) with empirical k values, because the samples may be contaminated by two factors: they may depend on each other, or/and they belong to different

distributions (Sections 4.2 and 4.3). Figure 3a shows such an example where the monthly wind maxima from a 5 year period at the site Jylex (black dots) suggest a type II distribution with $k < 0$ (the solid black curve, which was obtained through regression). The problem is caused by having samples from different mechanisms and therefore different initial distribution functions. In this plot, the lowest values are from summer, and they have a different distribution from those higher values from winter, where the larger temperature difference across latitudes results in more severe mid-latitude storms.

Davidson and Smith[12] suggested a graphical method of estimating k by the relation

$$E(U - u_0 | U > u_0) = [A - k(u_0 - u_{01})]/(1 + k) \tag{14}$$

where u_0 is an array of threshold, and u_{01} is the lowest of u_0. If the generalized Pareto assumption is correct, then the plot of the mean observed excess $E(U - u_0|U > u_0)$ versus $u_0 - u_{01}$ should follow a straight line with intercept $A/(1 + k)$ at $u_0 = u_{01}$ and slope of $-k/(1 + k)$. Thus, k and A can be obtained. This approach allows a systematical sensitivity check on the choice of u_0 and assures the robustness in the determination of k. Furthermore, from the previously described least-squares fit, the statistical uncertainty of the determination of the slope may be determined from the deviation of the actual E-values as a function of u_0 from the regression line and hence also in the resulting value of k. The details are given in Appendix B.

This approach was earlier used by Holmes and Moriarty,[13] and it is also used here to estimate the k-value for the five sites. The results from the longest record at Tystofte is shown in Figure 4, where Figure 4(a) shows the observed mean exceedance E in dependence of a series of threshold u_0 ranging from 11 to 27 ms^{-1}. The mid-latitude storms usually have a duration of several days. The measurements from Denmark suggest that individual extreme wind storms do not last more than 1 week. Here, a minimum time interval Δt of 7 days between two consecutive peaks over u_0 is used to

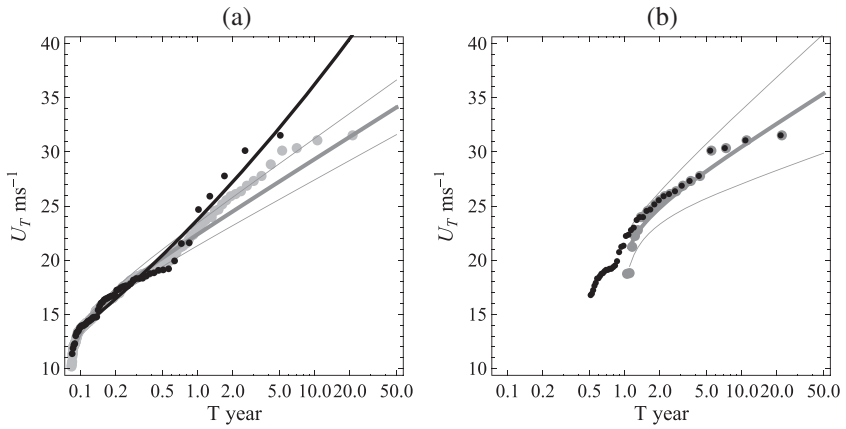

Figure 3. Distribution of U_T with T for Jylex site in log-linear scale. (a) Basis period $T_{BP} = 1$ month using a short time series of the first 5 years (black dots) and using the entire time series (gray dots); the Gumbel distribution with the 95% confidence intervals are shown in dashed lines for the gray dots. The solid, black curve is the regression line for the black dots. (b) Using the entire time series but with $T_{BP} = 6$ months (black dots) and $T_{BP} = 1$ year (gray dots). The Gumbel distribution with the 95% confidence intervals are shown in dashed lines for the gray dots.

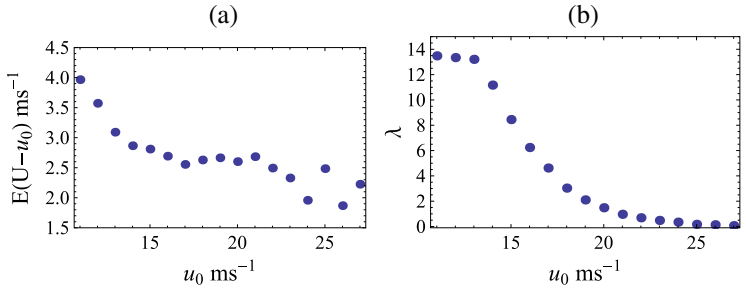

Figure 4. (a) Variation of the mean exceedance E as a function of threshold u_0 at Tystofte; (b) corresponding occurrence rate. $\Delta t = 7$ days was used.

Table III. The measure of GOF to the 95% confidence interval, $\max(\Delta F) - D_c$, for four k values, for the two distributions GEVD and GPD.

k	GEVD	GPD
0	−0.19	−0.03
0.004	−0.19	−0.02
0.055	−0.19	*0.04*
0.099	−0.18	*0.14*

The negative values pass the GOF test and the two positive values (in italic form) fail.

ensure them to be independent of each other. Ideally, in order to satisfy the condition that the samples are of identical distribution, the meteorological background of the identified peaks should be examined to ensure that they are of the same mechanism. Here, because of the simple extreme wind phenomenon, this procedure is simplified, and only the dates of the peaks are examined to confirm if they are all from the winter months. Seemingly, E decreases rapidly with u_0 for the first few numbers but slows down at higher u_0. At the same time, the exceedance rate λ decreases with u_0 and becomes smaller than once per year at u_0 about 21 ms^{-1} [Figure 4(b)]. The varying slope of E in different ranges of u_0 suggests that the regression coefficients for E versus $u_0 - u_{01}$, the slope and intercept, and accordingly, k and A, will depend on the choice of u_{01} and the range of u_0. To address this issue, three u_{01} in three different ranges of u_0 are examined:

(1) $u_{01} = 11$ ms^{-1} and u_0=[11, 21] ms^{-1}
(2) $u_{01} = 16$ ms^{-1} and u_0=[16, 27] ms^{-1}
(3) $u_{01} = 16$ ms^{-1} and u_0=[16, 21] ms^{-1}

In tests 1 and 3, the maximum of u_0 is chosen to be 21 ms^{-1} because λ becomes less than 1 when $u_0 > 21$ ms^{-1}. We denote the threshold where $\lambda \approx 1$ as u_{0h}. There have been discussions in the literature regarding which value to choose as u_{01}; some use a certain percentile such as 90 and 93%.[4,14] In the study by Holmes and Moriarty,[13] the minimum of the set of yearly maxima was used. In this study, we also use the minimum of the set of yearly maximum wind speed as u_0. Compared with the percentiles, the use of minimum annual maximum can better take into account of the year-to-year variation, which will be shown later in Section 4 to be highly important. On the contrary, using the percentile has the tendency to overlook years where the extreme winds are relatively lower. Thus, in tests 2 and 3, u_{01} is chosen to be 16 ms^{-1}, which is the minimum of the annual wind maxima, $\min(U^{max})$. The corresponding k and A for the three tests are accordingly obtained:

(1) $k = 0.099$, $A = 3.82$ ms^{-1}
(2) $k = 0.055$, $A = 3.22$ ms^{-1}
(3) $k = 0.004$, $A = 2.62$ ms^{-1}

In order to judge if the GEVD and GPD are good models for the samples, the goodness-of-fit (GOF) of the extreme value distribution is examined through the property $\max(|\Delta F|)$, which is the maximum difference between the predicted and measured $F(U)$. Depending on the number of observations and the level of significance, $\max(|\Delta F|)$ is examined whether it is smaller than a critical value D_c; if it is, then the distribution function is accepted. To the often used significant level 0.05, the corresponding confidence level is 0.95 and $D_c = 1.36/\sqrt{n}$, with n being the number of extreme wind events over u_0.[7] The values of $\max(\Delta F) - D_c$ for four k are listed in Table III, for both distributions, GEVD and GPD.

The GEVD has shown not to be sensitive to the k values from 0 to 0.099, as can be seen in Figure 5(a) and Table III, and equation (1) passed the GOF test for all four k values. On the other hand, the GPD is much more sensitive to the k effect [Figure 5(b) and Table III]. Here, equation (9) failed GOF for both $k = 0.055$ and $k = 0.099$ (the italic numbers in Table III), and using them with the corresponding A values has given descriptions of U_T versus T with the extreme wind samples outside the 95% intervals.

Similar sensitivity tests at the other four sites showed consistent results. Seemingly, for the sites used here, $k = 0$ is a good approximation. Thus, we carry on the analysis of the uncertainties in using PMM and POT with $k = 0$.

4.2. The PMM

One of the most obvious problems with a short time series is that it is difficult to collect extreme wind samples enough for a good distribution fit. Say, if we have 5 years of data, using AMM will give us five samples to make a fit; the uncertainty is significant according to equation (7). One popular fix is to use PMM with smaller T_{BP}, instead of using AMM. Thus, using $T_{BP} = 6$ months, we will have 10 samples to make the fit, and using $T_{BP} = 1$ month, we will have 60 samples! The risk of using a very small T_{BP} has been discussed at the beginning of Section 4.1 through Figure 3(a). The low extreme

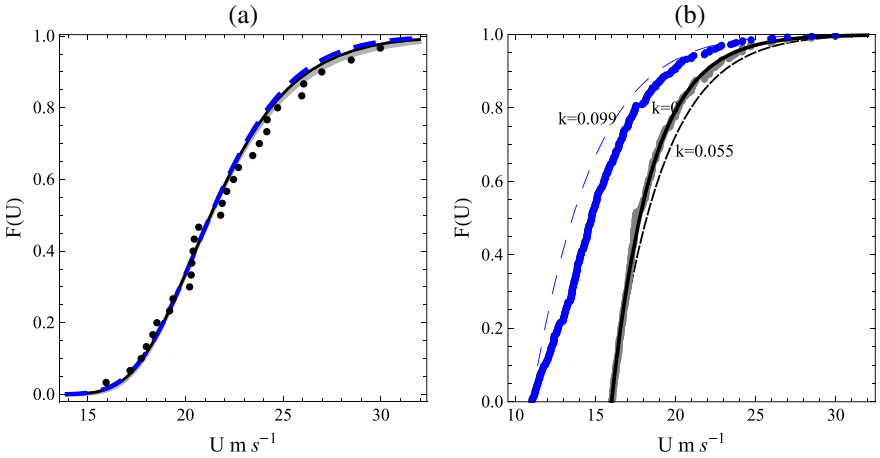

Figure 5. The distribution of extreme wind events $F(U)$ for (a) AMM: the curves correspond to $k = 0$ (thick gray), 0.055 (thin black) and 0.099 (dashed), respectively. (b) POT: the dots to the left are the measurements and the dashed curve is $F(U)$ for $u_0 = 11$ ms^{-1} and $k = 0.099$. The dots in the middle are the measurements and the solid curve is $F(U)$ for $u_0 = 16$ ms^{-1} and $k = 0$. The thin dashed curve to the right correspond to $u_0 = 16$ ms^{-1} and $k = 0.055$.

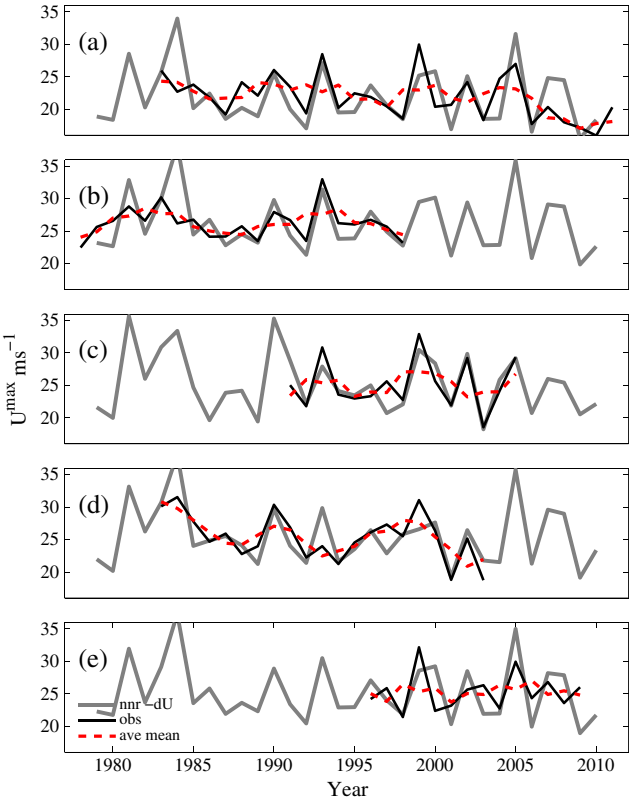

Figure 6. Yearly wind maxima at (a) Tystofte, (b) Sprogø, (c) Kegnæs, (d)Jylex and (e):Risø. Thick, gray curves: the yearly maximum geostrophic winds from the NCEP/NCAR re-analysis data at the closest grid point to the measurements, with the mean deficit of the measured and re-analysis values removed. Black, solid curve: measurements. Dashed curves: a 3 year running mean of the measured values.

winds from summer and high extreme winds from winter result in a negative shape factor k, which could lead to significant overestimation of U_T. The situation could be even more complicated in places where the extreme winds are from multiple mechanisms. The effect of a negative k becomes less dominating when the long time series of 24 years is used, owning to the increased number of strong wind samples [Figure 3(a), gray dots]. Even so, with $T_{BP} = 1$ month, the Gumbel estimate of U_T is out of the 95% confidence interval when $T > 2$ years. On the contrary, using $T_{BP} = 6$ months [Figure 3(b), black dots] and 1 year [Figure 3(b), gray dots] gives the estimate of U_T within the 95% confidence interval all the way through to 50 years. For the area we studied here, the extreme wind cases from the first or the second half of the year are both from the Atlantic lows in winter; it is reasonable to use $T_{BP} = 6$ months.

Another issue of using a very short time series (such as 5 years) is the representativity of this short period for a T-year extreme wind climate. In the extrapolation of the limited years of data to T years through a distribution function, we have assumed that the extreme wind climate is the same for the short period as for the T years. In Figure 6, the observed yearly wind maxima at the five sites are plotted (black, solid lines). In addition, the observed yearly wind maxima using a 3-year running mean are also plotted (dashed lines) for the observation period, in order to better see the long-term cyclic variation. From the dashed line, one could see an approximate cyclic period greater than 5 years. This cyclic variation possibly reflects an internal correlation of the yearly maximum on a scale of 8–10 years. Thus, using a period of 5 years or less has a very high chance missing either the ridges or the troughs of these curves, and the corresponding estimate of the T-year wind can be biased from using a time series with data length of decades.

This effect of using a limited period of data is shown in Figure 7, where U_{50} are calculated using AMM with data length, L, ranging from 3 years to the record length, N, for all five sites. The right-most dot in each subplot is the estimate from

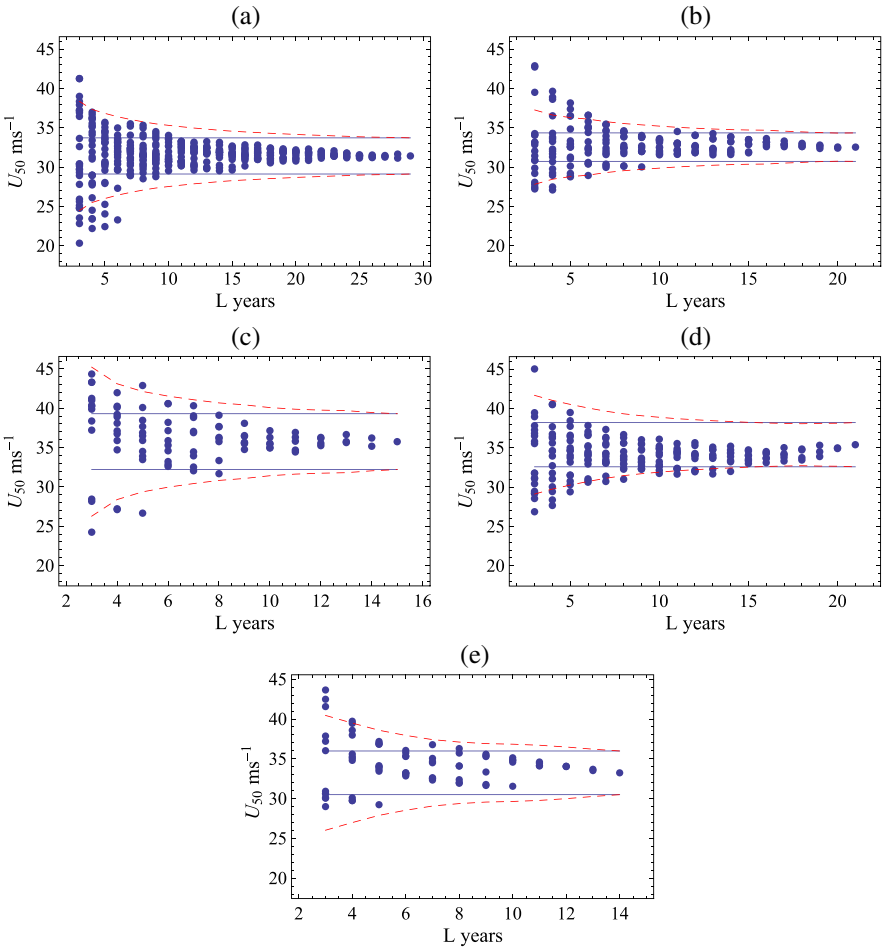

Figure 7. The 50 year wind U_{50} from AMM with data length L ranging from 3 years to the record length N, (a) Tystofte, (b) Sprogø, (c) Kegnæs, (d) Jylex and (e) Risø. In each subplot, the right-most dot is the estimate based on the entire time series, the solid straight lines show $\sigma(U_T)$ [equation (7)] centered at U_{50} from the entire time series, and the two dashed curves show $\sigma(U_T)$ at $L = 3$ to N, centered at U_{50} from the entire time series.

the entire time series where $L = N$, and the straight lines are $U_{50} \pm \sigma(U_{50})$ from equation (7) for the entire record length, $L = N$. The multiple values at each $L(< N)$ are calculated from L neighboring years. There are, thus, $N - L + 1$ segmentations for each L. The greater L, the fewer segmentations and therefore fewer values of U_{50}. Even though the spread of U_{50} is still of a good measure of the uncertainty and the spread becomes less at larger L, it seems to be approximately limited within $\pm\sigma(U_{50})$ for $L = N$ when $L > 10$ years. Clearly, the shorter the time series, the larger chance the estimation of U_{50} is biased. According to Figure 6, a data length longer than 10 years seems sufficient for these sites to include both troughs and ridges. However, if the long-term trend is present over several decades, it will still make a difference from which period a long term data of, e.g., 10 years, is used.

In Figure 7, $\pm\sigma(U_{50})$ at each L from 3 to N years is plotted as the dashed curves, centered at the least biased estimate, namely U_{50} with $L = N$. To read the uncertainty for each point in the plot, the dashed curves should be shifted vertically to be centered at that point.

4.3. The POT method

The samples of high wind events over a threshold u_0 are determined by the choice of u_0 and the time interval Δt that filters out the dependent cases. On the basis of the observation that a mid-latitude cyclone is of the duration of several days, setting $\Delta t = 7$ days could be considered as an appropriate start. The pre-condition of using POT is that u_0 is not too low. It has also been discussed in previous sections that low values of u_0 should be avoided because of several reasons. First, with a very low u_0, it is technically difficult to define 'peaks'. Second, using a too low u_0 can result in inclusion of wind speeds that are not due to the main extreme wind mechanism, which might modify the total distribution and lead to wrong estimate of U_T. The problem of using a too low u_0 is demonstrated with the Tystofte data in Figure 8, where $u_0 \ll \min(U^{max})$. The distribution for the lower winds seems different from that for the highest winds, similar to what Figure 3 shows. The estimation of U_T in Figure 8 is dominated by low wind samples. With $u_0 = 11$ ms^{-1}, $F(U)$ using the k and A values as used in Figure 8 did not pass the GOF test. The situation is similar at the other four sites. The POT method is often preferred to in comparison with PMM because the wind events are not bounded by a fixed period, and more events can be used. For this argument, it is reasonable, as we did in Section 4.1, not to set u_0 too high to avoid the occurrence rate $\lambda \ll 1$.

The results for the five sites are shown in Figure 9, where the range $\min(U^{max}) < u_0 < u_{0h}$ is marked in thick black curves on the x-axes. In Figure 9, we have used $\Delta t = 7$ days for all sites except for Kegnæs, for which $\Delta t = 10$ days was used. The choice of Δt is based on Figure 10, where the dependence of U_{50} on Δt are shown for a given u_0 at each site; the range of Δt from 1 day to 1 month is shown.

The interdependence of U_{50}, Δt and u_0 requires a series of sensitivity tests, and with the help of information as shown in Figures 4 and 9, one can find the reasonable pair of u_0 and Δt. However, the estimate of U_T is much more sensitive to u_0 than to Δt.

With a given u_0, it seems that the sensitivity of the results to Δt is vanishing for $\Delta t > 5$ days, although Figure 9 may look somehow different with other choice of u_0. At Tystofte, Sprogø, Jylex and Risø [Figure 10(a), (b), (d) and (e)], using $\Delta t \geqslant 7$ days seems to bring estimate of U_{50} to a stationary level. At Kegnæs, this stationary level is reached when $\Delta t \geqslant 10$ days. The different U_{50} corresponding to the first few Δt are mostly caused by the inclusion of lower wind samples.

Through the sensitivity tests for a range of Δt and u_0, the estimate of the 50 year wind U_{50} for these sites are obtained and shown in Figure 9. One can see that the U_{50} values are consistent when $\min(U^{max}) < u_0 < u_{0h}$ and the variation is within $\pm\sigma$. The corresponding values from AMM are shown in Figure 9 as thick gray lines, and they are in rather good agreement with the estimates from POT, with the difference in the range of $\pm\sigma(U_T)$.

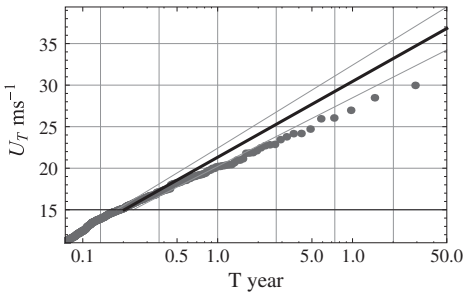

Figure 8. An example of using POT for U_T using a too low threshold u_0, from entire data record at Tystofte, $u_0 = 11$ ms^{-1}, $\Delta t = 7$ days, corresponding $\max(\Delta F) > D_c$.

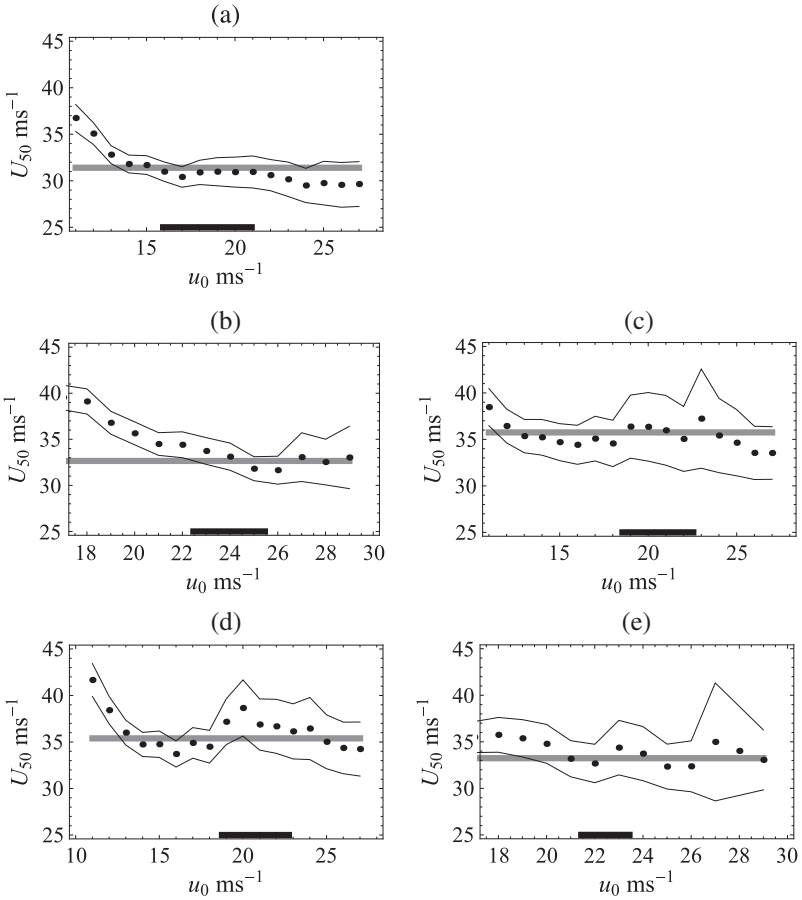

Figure 9. The 50 year wind U_{50} varying with threshold u_0. The two thin lines show $\pm\sigma(U_T)$ [equation (13)]. The thick gray curve shows the value from AMM using entire time series. The short black line marked on the *x*-axis on each plot shows the range starting with minimum of the annual maxima and ending with the threshold corresponding to occurrence rate of about one per year. (a) Tystofte ($\Delta t = 7$ days), (b) Sprogø ($\Delta t = 7$ days), (c) Kegnæs ($\Delta t = 10$ days), (d) Jylex ($\Delta t = 7$ days) and (e) Risø ($\Delta t = 7$ days).

Similar to PMM, the big concern with using short time series is still its climatological representativity of the extreme wind. In Figure 11, U_{50} using POT were shown as a function of data length L ranging from 3 years to the record length, in the same manner as Figure 7. The two plots, Figures 11 and 7, are very similar, except that Figure 11 corresponds to smaller uncertainty, given as $\sigma(U_T)$ from equation (13).

5. DISCUSSIONS

Short time series are often used in industry for the estimation of the extreme winds. This paper flags warnings and suggests possible solutions for such actions. It demonstrates the various sources to the uncertainties in the use of the two widely used distribution functions GEVD and GPD. This was done through the analysis of measurements from five stations in Denmark. The measurements here represent one single extreme wind mechanism, i.e., the Atlantic lows. This simplicity of the extreme wind characteristics rules out extra sources of uncertainty such as multiple extreme event mechanisms[6] and makes it easier for us to focus on a series of other factors that are more related to the use of short time series, embedded in the application of the two distribution functions.

There are no theoretical reasonings why GEVD and GPD should be used for extreme wind estimation, except that they have been empirically proven to be useful. The same is for the shape factor k in the two distribution functions. Whereas the non-zero k value can be truly related to the extreme wind distribution, it could also be a result of badly chosen u_0 or T_{BP} from using a short time series. The latter was demonstrated here by measurements that inappropriate choice of u_0 and Δt can result in mistaken k values and thereof wrong estimation of U_T. For these simple data from Denmark, $k = 0$ is a good approximation, and large values of $|k|$ were shown to be related to including low winds that are not from the Atlantic

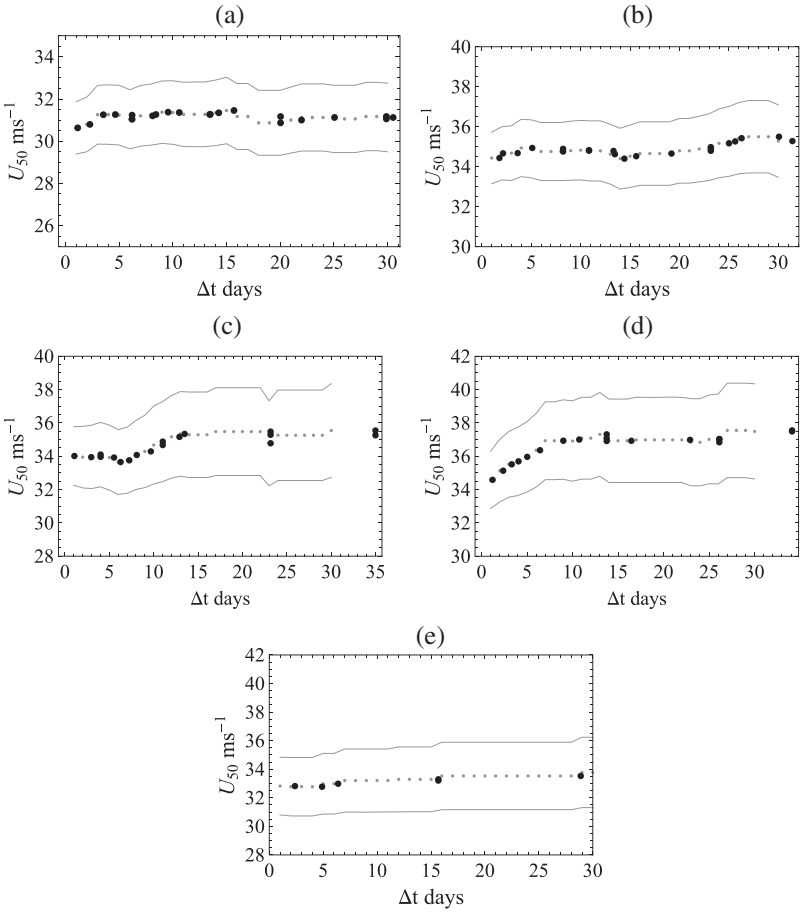

Figure 10. The 50 year wind U_{50} varying with Δt. Gray dots show U_{50} at required Δt, and black dots show U_{50} at the actual Δt. The thin gray curves are $\sigma(U_T)$ (a) Tystofte ($u_0 = 18$ ms^{-1}), (b) Sprogø ($u_0 = 22$ ms^{-1}), (c) Kegnæs ($u_0 = 20$ ms^{-1}), (d):Jylex ($u_0 = 19$ ms^{-1}) and (e) Risø ($u_0 = 22$ ms^{-1}).

lows when too small values of u_0 for GPD or too small basis period for GEVD were used. With $k = 0$, the two methods, PMM and POT, using GEVD and GPD, respectively, have provided consistent estimation of the return wind. This is not a surprise; equation (4) for PMM and equation (12) for POT are of the same form. For large samples, the values of A and B as in equation (12) are comparable with α^{-1} and β, as in equation (4).

For areas where the k effect may be important, it seems crucial how to choose the range of u_0 for obtaining the k and A values as demonstrated in the study by Holmes and Moriarty[13] as well as in the current study. Holmes and Moriarty[13] suggested to start the array of u_0 with a value about the minimum of the annual wind maxima, which seems to be a good choice for the data used here. In doing so, we pre-conditionally selected winds of the similar level to the annual maxima. We also recommend to end the array of u_0 at where λ turns 1 to ensure certain amount of samples. This choice is supported by the consistent estimates of U_T from PMM and POT using $k = 0$ and k-values obtained using this range of u_0 when the results start not to be sensitive to the choice of u_0 any more. This choice accordingly leaves very little chance for using time series of only a couple of years. At other times, when the k effect seems significant because multiple extreme wind mechanisms are involved, different approaches may be required, such as separating the samples according to their mechanisms.[6]

The first control of the applicability of a distribution function is to see whether it describes well the extreme wind samples through the GOF criteria. Here, we used the 95% level of confidence interval. Using this criteria did reject occasions where too low basis period and too low u_0 were used [Figures 3(a) and 8]. With this criteria, $k = 0$ seems a good approximation for these Danish sites. However, the 95% confidence interval is rather generous, e.g., using $k = 0.099$ for AMM also passed the GOF test. If we want a more strict criteria, higher significance levels can be used, e.g., $1.22/\sqrt{n}$ (significance level of 0.10, corresponding to confidence interval of 90%) and $1.07/\sqrt{n}$ (significance level of 0.20, corresponding to confidence interval of 80%).[7]

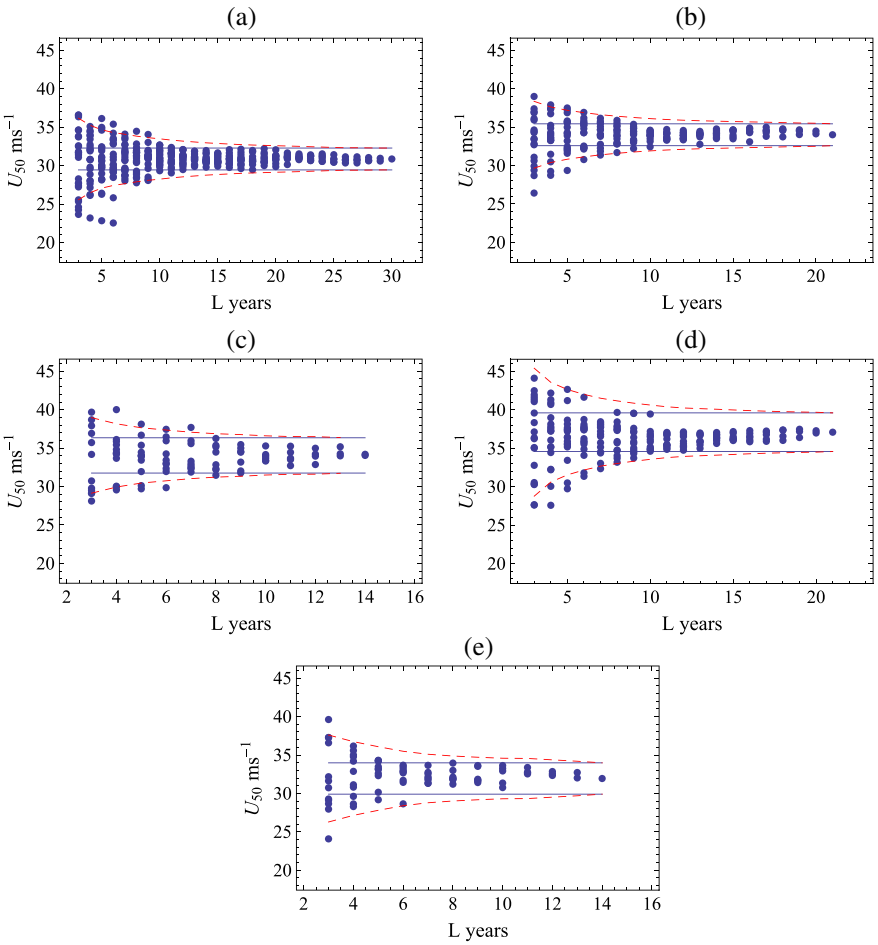

Figure 11. The 50 year wind U_{50} using POT with data length L from 3 years to the record length of N years. (a) Tystofte ($u_0 = 18$ ms^{-1}, $\Delta t = 7$ days), (b) Sprogø ($u_0 = 22$ ms^{-1}, $\Delta t = 7$ days), (c) Kegnæs ($u_0 = 20$ ms^{-1}, $\Delta t = 10$ days), (d) Jylex ($u_0 = 19$ ms^{-1}, $\Delta t = 7$ days), (e) Risø ($u_0 = 22$ ms^{-1}, $\Delta t = 7$ days). In each subplot, the right-most dot is the estimate based on the entire time series, the solid straight lines show $\sigma(U_T)$ [equation (13)] centered at U_{50} from the entire time series and the dashed curves show $\sigma(U_T)$ at $L = 3$ to L years, centered at U_{50} from the entire time series.

The relation of the confidence interval and sample size (equations 7 and 13) determines that, under the same significance level, the smaller sample size it is, the broader confidence interval it becomes. This leads to the fact that, when using a very short time series, it may happen that the distribution function passes the GOF criteria, but the estimate is biased. The combined effect of bias and fitting uncertainty of a range of sample size is shown with examples in Figures 7 and 11. One can see that at small L, the bias plus the fitting uncertainty give estimates that can be far away from the estimate at $L = N$.

The short time series has a high chance to fail representing the climatology of the extreme winds, regarding the inter-annual variation and long-term trend. Both AMM and POT give consistent results on the uncertainty related to using a limited number of years data (Figures 7 and 11), and apparently, the shorter time series has a larger chance for biased estimates and at the same time has much higher uncertainty. The spread of U_{50} is much more significant when using data shorter than 10 years. In Larsén *et al.*,[15] the long-term variability of the extreme wind was studied using wind measurements together with geostrophic winds calculated with outputs from a number of global and mesoscale models. Different models have shown consistent long-term extreme wind variability, which agrees with the measurements. The consistency between point measurements and coarse resolution modeled data is owing to the fact that the extreme winds in the studied region are controlled by the synoptic storms related to the North Atlantic oscillation. Under this concept, the geostrophic winds from 1979 to 2010 at the corresponding National Centers for Environmental Prediction (NCEP)/National Center for Atmospheric Research (NCAR) re-analysis grid points to the five sites were calculated here. In order to better show the relationship between the measured and modeled data, for their overlapping period, the deficit of the measured and modeled mean annual maxima was removed from the modeled geostrophic wind and the remaining is plotted in Figure 6 as the thick

gray curves for each site. Even though it is a great advantage that the reanalysis data are globally available for decades, because of the coarse temporal and spatial resolutions that do not match with the measurements, these data cannot be used directly for site-specific extreme wind estimation.[16] However, the trend and the interannual extreme wind variability are of large scale, and they seem to be well represented in coarse resolution data (Figure 6).[15] This implies that the long-term re-analysis products can be used to provide climatological information into the short time series.

With the long-term effect taken into account, Larsén and Mann[16] calculated the extreme wind from the NCEP/NCAR reanalysis data through the geostrophic wind; through microscale modeling, the re-analysis data are downscaled to site-specific extreme wind values. In this approach, the mesoscale variability is completely neglected, which could be a serious flaw in places where the extreme winds are of mesoscale origins. Even the computationally costly mesoscale model runs are limited in resolving the range of wind variabilities required for extreme wind estimation because of the spatial and temporal smoothing effect embedded in the numerical computations.[5] A statistical model was developed by Larsén *et al*.[5] to feed in the missing variabilities to the modeled time series through the power spectrum, in order to improve the estimation of the extreme wind from the modeled output. The aforementioned approaches for taking into account of the extreme wind climate are rather computationally costly. A simpler solution could be to use the spectral correction approach directly to the reanalysis data.[17] Anastasiades and McSharry[18] recently developed a mathematical spectral approach to add in seasonal and yearly wind variation from re-analysis data to the limited measurements, and they have shown good results. Seemingly, these most up-to-date studies all suggest a combination of modeled data with measurements if they are short, through physics or statistical approaches.

6. CONCLUSIONS

The conclusions from the study can be summarized as:

(1) This study can be used as a guideline for using GEVD and GPD with wind time series of limited length to understand the uncertainties in the extreme wind estimation.
(2) With reasonable choice of relevant parameters, PMM and POT give consistent estimates of the return winds.
(3) For POT, the choices of u_0, Δt and k are interrelated. Δt can be obtained on the basis of the physics of the extreme wind events, and the sensitivity can be examined. It is recommended to use u_0 within the approximate range of the minimum of the annual wind maximum and where the occurrence rate becomes less than 1. For a simplified calculation, it is recommended to use $k = 0$. For cases with significant k values, it is needed to find out the causes, and special approaches might be needed to estimate the 50 year wind.
(4) The lack of climatological representativity is a major source of uncertainty, and the information of climatological variability can be considered to be extracted from global or mesoscale models.

ACKNOWLEDGEMENTS

This study is partly supported by Danish grant MesoExtremes 2009-1-10240 and partly by the project of Wind Atlas of South Africa (WASA). The NCEP/NCAR data are provided by NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, from their website at http://www.cdc.noaa.giv/.

APPENDIX A

Relationship between GEVD and GPD for the same wind climate

For the same wind climate, the shape factor k of GEVD has in fact the same value as that of GPD. This is so since the PMM and POT distributions are strictly related as, e.g., given by Abild (Section 6.6.1).[7] Abild[7] uses that the accumulative distribution for the periodical maximum speed U is in fact the probability that, within the base period T_{BP}, U is not exceeded. The relation is then

$$F_{PMM}(U;T_{BP}) = \exp\left(-T_{BP}\lambda_0(1-F_{POT}(U;u_0))\right) \tag{A1}$$

or with the GEVD substituted for the PMM distribution, and GPD for POT

$$\begin{aligned} F_{GEVD}(U,T_{BP}) &= \exp\left(-T_{BP}\lambda_0\left(1-\frac{k(U-u_0)}{A}\right)^{1/k}\right), \quad k\neq 0 \\ &= \exp\left(-T_{BP}\lambda_0\exp\left(\frac{U-u_0}{A}\right)\right), \quad k=0 \end{aligned} \tag{A2}$$

This is identical to equations (1) and (2) with:

- the same k value
- A and u_0 related to the GEVD parameters as

$$\begin{aligned} k \neq 0: \ \alpha &= \frac{(T_{BP}\lambda_0)^k}{A}; \ \ \beta = u_0 + \frac{A - \frac{1}{\alpha}}{k} \\ k = 0: \ \alpha &= 1/A; \ \ \beta = u_0 + A\ln(T_{BP}\lambda_0) \end{aligned} \tag{A3}$$

APPENDIX B

Statistical uncertainty of the shape factor *k* from the least square regression procedure

Let the mean speed exceedance be determined for a number (M) of suitably spaced speed threshold values ($E_j, u_{0,j}$) from the time series to be analyzed. Then, the slope S of the regression line and the statistical uncertainty of it, $\sigma(S)$, can be obtained as follows:

First, for convenience, we introduce the shifted variables:

$$\begin{aligned} \Delta u_{0,j} &= u_{0,j} - \overline{u}_0 \\ \Delta E_j &= E_j - \overline{E} \end{aligned} \tag{B1}$$

where

$$\begin{aligned} \overline{u}_0 &= \frac{1}{M}\sum_{j=1}^{M} u_{0,j} \\ \overline{E} &= \frac{1}{M}\sum_{j=1}^{M} E_j \end{aligned} \tag{B2}$$

Then the regression procedure proceeds as

$$S = R_2/Q_{22} \tag{B3}$$

where

$$R_2 = \sum_{j=1}^{M} \Delta E_j \Delta u_{0,j}, \ \ Q_{22} = \sum_{j=1}^{M} (\Delta u_{0,j})^2$$

and

$$\sigma^2(S) = \frac{\chi^2}{Q_{22}(M-2)} \tag{B4}$$

where

$$\chi^2 = \sum_{j=1}^{M} (\overline{E} + S\Delta u_{0,j} - E_j)^2.$$

Hence, the shape parameter k and the statistical uncertainty of it, $\sigma(k)$, are found as

$$\begin{aligned} k &= \frac{-S}{1-S} \\ \sigma(k) &= \frac{\sigma(S)}{(1-S)^2} \end{aligned} \tag{B5}$$

REFERENCES

1. Eurocode. Eurocode 1, Basis of design and actions on structure – Parts 2 – 4: Actions on structure – Wind actions. *Technical Report*, European Committee for Standardization, Rue de Stassart, Brussels, 1995.
2. Palutikof JP, Brabson BB, Lister DH, Adcock ST. A review of methods to calculate extreme wind speeds. *Meteorological Applications* 1999; **6**: 119–132.
3. Alexandersson H, Schmith T, Iden K, Tuomenvirta H. Long-term variations of the storm climate over NW Europe. *The Global Atmosphere and Ocean System* 1998; **6**: 97–120.
4. Vinoth J, Young IR. Global estimates of extreme wind speed and wave height. *Journal of Climate* 2011; **24**: 1647–1665.
5. Larsén XG, Ott S, Badger J, Hahmann AH, Mann J. Recipes for correcting the impact of effective mesoscale resolution on the estimation of extreme winds. *Journal of Applied Meteorology and Climatology* 2012; **51**(3): 521–533. DOI: 10.1175/JAMC-D-11-090.1.
6. Cook N, Harris R, Whiting R. Extreme wind speeds in mixed climates revisited. *Journal of Wind Engineering and Industrial Aerodynamics* 2003; **91**: 403–422.
7. Abild J. Application of the wind atlas method to extremes of wind climatology. *Technical Report Risoe-R-722(EN)*, Risø National Laboratory, Roskilde, Denmark, 1994.
8. Hosking JRM. Estimation of the generalized extreme value distribution by the method of probability-weighted moments. *Technometrics* 1985; **27**: 251–261.
9. Mann J, Kristensen L, Jensen NO. Uncertainties of extreme winds, spectra and coherences. In *Bridge Aerodynamics, ISBN 9054109610*, Larsen, Esdahl (eds). Balkema: Rotterdam, 1998; 49–56.
10. Kite GW. Confidence limits for design events. *Water Resources Research* 1975; **11**: 48–53.
11. Cook N. Towards better estimation of wind speeds. *Journal of Wind Engineering and Industrial Aerodynamics* 1982; **9**: 295–323.
12. Davison A, Smith R. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B* 1990; **52**: 393–442.
13. Holmes J, Moriarty W. Application of the generalized pareto distribution to extreme value analysis in wind engineering. *Journal of Wind Engineering and Industrial Aerodynamics* 1999; **83**: 1–10.
14. Caires S, Sterl A. 100-year return value estimates for ocean wind speed and significant wave height from the ERA-40 data. *Journal of Climate* 2005; **18**: 1032–1048.
15. Larsén XG, Mann J, Göttel H, Jacob D. Wind climate and extreme winds from the regional climate model REMO, *Scientific Proceedings (on line). European Wind Energy Conference and Exhibition*, Brussels, 31 March - 3 April, 2008; 58–62.
16. Larsén XG, Mann J. Extreme winds from the NCEP/NCAR reanalysis data. *Wind Energy* 2009; **12**: 556–573. DOI: 10.1002/we.318.
17. Larsén XG, Kruger A, Badger J, Jørgensen HE. Extreme wind atlases of South Africa from global reanalysis data, *Proceedings of the 6th European and African Conference on Wind Engineering*, Cambridge, UK, July, 2013.
18. Anastasiades G, McSharry P. Extreme value analysis for estimating 50-year return wind speeds from reanalysis data. *Wind Energy, published online* 2013. DOI: 10.1002/we.1630.